ABSTRACT

Left truncation often arises when patient information, such as time of diagnosis, is gathered retrospectively. In some cases, the distribution function, say $G(x)$, of left-truncated variables can be parameterized as $G(x;\theta)$, where $\theta \in \Theta \subset R^q$, and θ is a q-dimensional vector. Under semiparametric transformation models, we demonstrated that the approach of Chen et al. (2002) can be used to analyze this type of data. The asymptotic properties of the proposed estimators are derived. A simulation study is conducted to investigate the performance of the proposed estimators.

Key Words: Left truncation, conditional likelihood, length-biased

1. Introduction

Randomly truncated data occur in many areas like astronomy, economics (see Woodroofe (1985), Fiegelson and Babu (1992)), epidemiology and biometry (see Keiding et al. (1987) and Lagakos et al. (1988)). Consider the following applications. For instance, in epidemiology, a prevalent cohort is defined as a group of diseased individuals who are recruited for a prospective study. The main interest of a research project is to study the natural history of the disease for individuals who developed the disease during the calendar time period (τ_0, τ), $\tau_0 < \tau$. Consider the sampling under which all of the individuals in the area who have experienced the first event (e.g. HIV infection) between τ_0 and τ and have not experienced a second event (e.g. diagonsed with AIDS) are recruited at the time τ for a prospective follow-up study. Let T_s denote the initial time of the first event and T be the time from T_s to the second event. Let V denote the time from T_s to τ. Hence, random left truncation occurs since one observes pair (V, T) only if $T \geq V$.

For non-infectious diseases, it may be reasonable to assume that the incidence process (denoted by $N(t)$) of a disease is regular, i.e. $P(N(t+\Delta t) - N(t) > 1) = o(\Delta t)$ as $\Delta t \to 0$, the unordered incidence times on a given interval $[u, u+k]$ are independent, and the intensity function of the process is constant. Under this case, the distribution of V, say $G(x) = P(V \leq x)$, can be parameterized as $G(x) = x/\theta$ for $x \in (0, \theta)$, the so-called stationarity assumption or length-biased sampling (see Wang (1991) and Asgharian et al. (2002)). For a new disease, however, one might prefer to parameterize G so that the parameterization reflects the growth of the disease over time, i.e. $G(x) = G(x; \theta)$, where $\theta \in \Theta \subset R^q$, and θ is a q-dimensional vector. Assume for each individual, data is available on some covariates Z. It is important to investigate the association between these covariates and survival rate.

Following the notations in Example, let T and V denote the lifetime and truncation time, respectively. Let $Z = [Z_1, \ldots, Z_p]^T$ represent a $p \times 1$ vector of covariates. Assume that given Z, T and V are independent of each other. For left-truncated data, one can observe nothing if $T < V$ and observe (T, V, Z) if $T \geq V$. In this article, we consider the following transformation model:

$$S(t|Z) = g\{h(t) + \beta^T Z\}, \tag{1.1}$$

where $S(t|Z) = P(T > t|Z)$ is the survival function of T given Z, the continuous, strictly

decreasing link function $g(\cdot)$ is given or specified up to a finite-dimensional parameter, $h(\cdot)$ is a completely unspecified strictly increasing function, and β is a $p \times 1$ vector of unknown regression coefficients. Note that when $g(\cdot) = \exp\{-\exp(\cdot)\}$, (1.1) gives the Cox proportional hazard model (Cox, 1972), and when $g(\cdot) = 1/\{1+\exp(\cdot)\}$ it corresponds to the proportional odds model (Bennett (1983); Murphy et al.(1997); Ying and Prentice (1999)). Furthermore, model (1.1) has an equivalent form (see Cheng et al. (1995))

$$h(T) = -\beta^T Z + \epsilon,$$

where the distribution of the error ϵ is $P(\epsilon \leq x) = F_\epsilon(x) = 1 - g(x)$. This can be easily justified by writing $S(t|Z) = P(h(T) > h(t)|Z) = P(-\beta^T Z + \epsilon > h(t)|Z)$.

When $g(\cdot)$ is completely specified, Chen et al. (2002) proposed an estimation procedure for the analysis of right-censored data. The procedure proposed by Chen et al. (2002) is easily implemented numerically and the estimator is the same as the Cox partial likelihood estimator in the case of the proportional hazards model. Shen (2012) extended Chen et al.'s approach to left-truncated and right -censored data. In this article, when the distribution function of V can be parameterized as $G(x;\theta)$, we demonstrated that the approach of Chen et al. (2002) can be used to obtain consistent estimators of β and $h(\cdot)$. The asymptotic properties of the proposed estimators are derived. In Section 3, a simulation study is conducted to investigate the performance of the proposed estimators.

2. The Proposed Estimators

Let $F(t|Z) = P(T \leq t|Z)$ denote the cumulative distribution function of T given Z. Suppose that the left and right endpoints of T are independent of Z. Let a_F and b_F denote the left and right endpoints of F, and similarly, define (a_G, b_G) as the left and right endpoint of V. We assume that both T and V are continuous. Furthermore, for identifiabilities of $F(t|Z)$, we assume that $a_G \leq a_F$ and $b_G \leq b_F$. First, we consider the case when there is no assumption on G. Let (T_i, V_i, Z_i) $(i = 1, \ldots, n)$ be the observed truncated sample. Let $Y_i(t) = I_{[V_i \leq t \leq T_i]}$ and $N_i(t) = I_{[T_i \leq t]}$. Let $p(Z_i) = P(V \leq T|Z_i)$. Note that $E[Y_i(t)|Z_i] = P(V_i \leq t \leq T_i|Z_i) = p(Z_i)^{-1}P(V \leq t|Z_i)P(T \geq t|Z_i)$ and $E[N_i(t)|Z_i] = p(Z_i)^{-1}P(V \leq T, T \leq t|Z_i)$. Let $\mathcal{F}(t)$ denote the filtration generated by

$$\sigma\{V_i, Z_i, Y_i(x), I_{[V_i \leq T_i]}, I_{[V_i < T_i \leq t]}, x \leq t; i = 1, \ldots, n\},$$

where $\sigma\{A\}$ denotes the complete σ-field generated by A. Let $\Lambda_\epsilon(\cdot)$ denote the cumulative hazard functions of ϵ, and $h_0(\cdot)$ and β_0 are the true values of $h(\cdot)$ and β, respectively. Under model (1.1), since

$$E[dN_i(t)|Z_i, \mathcal{F}_{t-}] = P(t \le T_i < t + dt|Z_i, \mathcal{F}_{t-}) = P(t \le T_i < t + dt, V_i < t|Z_i, \mathcal{F}_{t-})$$

$$= Y_i(t)d\Lambda_\epsilon(\beta^T Z_i + h_0(t)),$$

it follows that $M_i(t)$ is a martingale process with respect to $\mathcal{F}_t$. Using the approach of Chen et al. (2002), Shen (2012) considered the following two estimating equations:

$$U(\beta, h) = \sum_{i=1}^{n} \int_0^\infty Z_i[dN_i(t) - Y_i(t)d\Lambda_\epsilon(\beta^T Z_i + h(t))] = 0, \tag{2.1}$$

and

$$\sum_{i=1}^{n}[dN_i(t) - Y_i(t)d\Lambda_\epsilon(\beta^T Z_i + h(t))] = 0, \tag{2.2}$$

Step 0: Choose an initial value of β, denoted by $\hat{\beta}^{(0)}$.

Step 1: Let $t_1 < t_2 < \cdots < t_n$ denote the ordered failure time . Based on (2.2), we obtain $\hat{h}^{(0)}(t_1; \hat{\beta}^{(0)})$ by solving

$$\sum_{i=1}^{n} Y_i(t_1)\Lambda(\beta^T Z_i + h(t_1)) = 1,$$

with $\beta = \hat{\beta}^{(0)}$. Then, obtain $\hat{h}(t_k)$ for $k = 2, \ldots, n$, one-by-one by solving the equation

$$\sum_{i=1}^{n} Y_i(t_k)\Lambda(\beta^T Z_i + h(t_k)) = 1 + \sum_{i=1}^{n} Y_i(t_k)\Lambda(\beta^t Z_i + h(t_k-)),$$

with $\beta = \hat{\beta}^{(0)}$

Step 2: Obtain a new estimate of β by solving (2.1) with $h(t_k) = \hat{h}^{(0)}(t_k; \hat{\beta}^{(0)})$.

Step 3: Set $\hat{\beta}^{(0)}$ to be the estimate obtained in Step 2 and repeat Steps 1 and 2 until prescribed convergence criteria are met.

Let $(\hat{\beta}_n, \hat{h}_n)$ denote the solution of (2.1) and (2.2). It is then clear that $\hat{h}_n(\cdot) \in \mathcal{H}$. Note that $\hat{h}_n(\cdot)$ is a step function in t that rises at the distinct jump points of $\{I_{[T_i \le t]}; i = 1, \ldots, n\}$.

Similar to Proposition of Chen et al. (2002), under suitable regularity conditions, it follows that $n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0) \to N(0, \Sigma_{\hat{\beta}_n})$ in distribution, as $n \to \infty$, where $\Sigma_{\hat{\beta}_n} = \Sigma_2^{-1}\Sigma_1(\Sigma_2^{-1})^T$

$$\Sigma_1 = E\bigg[\int_0^\infty [Z_1 - \mu_z(t;\beta_0)]^{\otimes 2}\lambda_\epsilon(h_0(t) + \beta_0^T)Y_1(t)]dh_0(t)\bigg],$$

$$\Sigma_2 = E\bigg[\int_0^\infty [Z_1 - \mu_z(t;\beta_0)]Z_1^T\dot{\lambda}_\epsilon(h_0(t) + \beta_0^T)Y_1(t)]dh_0(t)\bigg],$$

where $x^{\otimes 2} = xx^T$,

$$\mu_z(t;\beta_0) = \frac{E[Z_1\lambda_\epsilon(h_0(T_1) + \beta_0^T Z_1)Y_1(t)B(t;T_1)]}{E[\lambda_\epsilon(h_0(t) + \beta_0^T Z_1)Y_1(t)]},$$

where $\lambda_\epsilon(\cdot)$ is the hazard functions of ϵ, $\dot{\lambda}_\epsilon(x) = d\lambda_\epsilon(x)/dx$ and

$$B(t,s) = \exp\bigg(\int_s^t \frac{E[\dot{\lambda}_\epsilon(h_0(x) + \beta_0^T Z_1)Y_1(x)]}{E[\lambda_\epsilon(h_0(x) + \beta_0^T Z_1)Y_1(x)}dh_0(x)\bigg).$$

Note that Σ_1 and Σ_2 can be consistently estimated by

$$\hat{\Sigma}_{1n} = n^{-1}\sum_{i=1}^n \int_0^\infty [Z_i - \bar{Z}(t;\hat{\beta}_n)]^{\otimes 2}\lambda(\hat{\beta}_n^T Z_i + \hat{h}_n(t))Y_i(t)d\hat{h}_n(t),$$

and

$$\hat{\Sigma}_{2n} = n^{-1}\sum_{i=1}^n \int_0^\infty [Z_i - \bar{Z}(t;\hat{\beta}_n)]Z_i^T\dot{\lambda}_\epsilon(\hat{\beta}_n^T Z_i + \hat{h}_n(t))Y_i(t)d\hat{h}_n(t),$$

respectively, where

$$\bar{Z}(t;\hat{\beta}_n) = \sum_{i=1}^n \frac{Z_i\lambda_\epsilon(\hat{\beta}_n^T Z_i + \hat{h}_n(t))Y_i(t)\hat{B}_n(t,T_i)}{\sum_{i=1}^n \lambda(\hat{\beta}_n^T Z_i + \hat{h}_n(t))Y_i(t)},$$

$$\hat{B}_n(t,s) = \exp\bigg(\int_s^t \frac{\sum_{i=1}^n \dot{\lambda}(\hat{\beta}_n^T Z_i + \hat{h}_n(x))Y_i(x)}{\sum_{i=1}^n \lambda(\hat{\beta}_n^T Z_i + \hat{h}_n(x))Y_i(x)}d\hat{h}_n(x)\bigg).$$

Hence, a consistent estimator of $\Sigma_{\hat{\beta}_n}$ is given by $\hat{\Sigma}_{\hat{\beta}_n} = \hat{\Sigma}_{2n}^{-1}\hat{\Sigma}_{1n}(\hat{\Sigma}_{2n}^{-1})^T$.

Now, under the assumption that $V \sim G(x;\theta)$, we shall propose an alternative estimator which incorporates the available information on the truncation distribution. Let θ_0 be the true value of θ. Let $\mathcal{F}_G(t)$ denote the filtration generated by

$$\sigma\{Z_i, \tilde{Y}_i(x;\theta_0), I_{[T_i \le x]}, ; x \le t\},$$

where $\tilde{Y}_i(x;\theta_0) = I_{[T_i \geq x]} G(x;\theta_0)/G(T_i;\theta_0)$. Define

$$\tilde{M}_i(t) = N_i(t) - \int_0^t \tilde{Y}_i(x;\theta_0) d\Lambda_\epsilon(\beta_0^T Z_i + h_0(t)) \quad (i = 1, \ldots, n).$$

Notice that under the assumption that $V \sim G(x;\theta_0)$, we have

$$E[I_{[V_i < x \leq T_i]} | T_i] = I_{[T_i \geq x]} P(V < x | V \leq T_i)$$

$$= I_{[T_i \geq x]} G(x;\theta_0)/G(T_i;\theta_0).$$

Hence, given T_i, the conditional distribution of $I_{[V_i < x \leq T_i]}$ is the same as that of $\tilde{Y}_i(x;\theta_0)$. It follows that

$$P(dN_i(t) = 1 | \mathcal{F}_w(t-)) = \tilde{Y}_i(t;\theta_0) d\Lambda_\epsilon(\beta_0^T Z_i + h_0(t)).$$

Hence, under model (1.1) and the assumption that $V \sim G(x;\theta_0)$, $\tilde{M}_i(t)$ is a martingale with respect to $\mathcal{F}_G(t)$. Based on the arguments above, given $G(\cdot;\theta_0)$, the estimators, say $(\tilde{\beta}_n(\theta_0), \tilde{h}(\cdot, \tilde{\beta}_n(\theta_0))$ can be obtained by simultaneously solving the following two equations:

$$\tilde{U}_{10}(\beta, h(\cdot)) = \sum_{i=1}^n \int_0^\infty Z_i [dN_i(t) - \tilde{Y}_i(t;\theta_0) d\Lambda_\epsilon(\beta^T Z_i + h(t))] = 0, \tag{2.3}$$

and

$$\tilde{U}_{20}(\beta, h(\cdot)) = \sum_{i=1}^n [dN_i(t) - \tilde{Y}_i(t;\theta_0) d\Lambda_\epsilon(\beta^T Z_i + h(t))] = 0. \tag{2.4}$$

For the special case of length-biased data and Cox model, i.e. $G(x;\theta) = x/\theta$ and $\lambda_\epsilon = \exp(t)$, it then follows from (2.3) and (2.4) that $\tilde{Y}_i(t;\theta_0) = \tilde{Y}_i(t) = I_{[T_i \geq t]} t/T_i$ and the estimator $\tilde{\beta}_n(\theta_0)$ satisfies the following equation:

$$\sum_{i=1}^n \int_0^\infty \left\{ Z_i - \frac{\sum_{j=1}^n Z_j \tilde{Y}_j(t) \exp(\beta^T Z_j)}{\sum_{j=1}^n \tilde{Y}_j(t) \exp(\beta^T Z_j)} \right\} dN_i(t) = 0,$$

which is precisely the Cox partial likelihood score equation for length-biased data (Wang (1996)).

Next, we consider the estimation of θ. Using the approach of Wang (1989), we consider the conditional likelihood of V_i's given T_i's as follows:

$$L_c(\theta) = \prod_{i=1}^n \frac{g(V_i;\theta)}{G(T_i;\theta)},$$

where $g(x;\theta)$ is the probability density function of $G(x;\theta)$. Let $\hat{\theta}_n$ be the maximizer of the estimated likelihood $L_c(\theta)$. Hence, given $\hat{\theta}_n$, alternative estimators of β_0 and $h_0(t)$, denoted by $(\tilde{\beta}_n, \tilde{h}_n(\cdot, \tilde{\beta}_n))$, can be obtained by simultaneously solving the following two equations:

$$\tilde{U}_{1n}(\beta, h(\cdot)) = \sum_{i=1}^{n} \int_0^{\infty} Z_i[dN_i(t) - \tilde{Y}_i(t;\hat{\theta}_n)d\Lambda_\epsilon(\beta^T Z_i + h(t))] = 0, \tag{2.5}$$

and

$$\tilde{U}_{2n}(\beta, h(\cdot)) = \sum_{i=1}^{n} [dN_i(t) - \tilde{Y}_i(t;\hat{\theta}_n)d\Lambda_\epsilon(\beta^T Z_i + h(t))] = 0. \tag{2.6}$$

For the special case of the Cox model, it then follows from (2.5) and (2.6) that the estimator $\tilde{\beta}_n$ satisfies the following equation:

$$\sum_{i=1}^{n} \int_0^{\infty} \left\{ Z_i - \frac{\sum_{j=1}^{n} Z_j \tilde{Y}_j(t;\hat{\theta}_n)\exp(\beta^T Z_j)}{\sum_{j=1}^{n} \tilde{Y}_j(t;\hat{\theta}_n)\exp(\beta^T Z_j)} \right\} dN_i(t) = 0,$$

which is precisely the Cox partial likelihood score equation for length-biased data (Wang (1996)).

Next, the following theorem can derived based on the argument of Chen et al. (2002, see Appendix)

Theorem 1. Under regularity conditions (Fleming and Harrington, 1991) for ensuring the central limit theorem for counting process martingales holds and assuming that (a) $G(x;\theta)$ is continuous in x for each $\theta \in \Theta$. (b) $\hat{\theta}_n \to \theta_0$ implies $G(x;\hat{\theta}_n) \to G(x;\theta_0)$ for each x and (c) Z_i is bounded, then (i) $n^{1/2}(\tilde{\beta}_n - \beta_0)$ converges in distribution to $N(0, \tilde{\Sigma}_{\tilde{\beta}_n})$, where $\tilde{\Sigma}_{\tilde{\beta}_n} = \tilde{\Sigma}_2^{-1}\tilde{\Sigma}_1(\tilde{\Sigma}_2^{-1})^T$, $\tilde{\Sigma}_1$ and $\tilde{\Sigma}_2$ are given in (2.7) and (2.8), respectively, and (ii) $\tilde{h}(\cdot, \tilde{\beta}_n)$ is consistent under the metric $d(\cdot,\cdot)$, where for any two nondecreasing functions h_1 and h_2 on $[0,\infty)$ such that $h_1(0) = h_2(0) = -\infty$, $d(h_1, h_2) = \sup(|\exp\{h_1(t)\} - \exp\{h_2(t)\}| : t \in [0,\infty))$

$$\tilde{\Sigma}_1 = E\left[\int_0^{\infty} [Z_1 - \tilde{\mu}_z(t;\beta_0)]^{\otimes 2} \lambda_\epsilon(h_0(t) + \beta_0^T)\tilde{Y}_1(t;\theta_0)]dh_0(t)\right], \tag{2.7}$$

where

$$\tilde{\mu}_z(t;\beta_0) = \frac{E[Z_1\lambda_\epsilon(h_0(T_1) + \beta_0^T Z_1)\tilde{Y}_1(t;\theta_0)\tilde{B}(t;T_1)]}{E[\lambda_\epsilon(h_0(t) + \beta_0^T Z_1)\tilde{Y}_1(t;\theta_0)]]}.$$

Next, similar to Step A4 of Chen et al. (2002), we have

$$\tilde{U}_{10}(\beta_0, \tilde{h}(\cdot, \beta_0)) = \sum_{i=1}^{n} \int_0^\infty Z_i d\tilde{M}_i(t) - \sum_{i=1}^{n} Z_i[\Lambda_\epsilon(\beta_0^T Z_i + \tilde{h}(T_i, \beta_0)) - \Lambda_\epsilon(\beta_0^T Z_i + h_0(T_i))]$$

$$= \sum_{i=1}^{n} \int_0^\infty Z_i d\tilde{M}_i(t) - \sum_{i=1}^{n} \frac{Z_i \lambda_\epsilon(\beta_0^T Z_i + h_0(T_i))}{\lambda^*(h_0(T_i))}[\Lambda^*(\tilde{h}(T_i, \beta_0)) - \Lambda^*(h_0(T_i))] + o_p(n^{1/2})$$

$$= \sum_{i=1}^{n} \int_0^\infty [Z_i - \tilde{\mu}_z(t)] d\tilde{M}_i(t) + o_p(n^{1/2}).$$

$$\tilde{\Sigma}_2 = E\left[\int_0^\infty [Z_1 - \tilde{\mu}_z(t; \beta_0)] Z_1^T \dot{\lambda}_\epsilon(h_0(t) + \beta_0^T)\tilde{Y}_1(t; \theta_0)] dh_0(t)\right]. \tag{2.8}$$

Note that $\tilde{\Sigma}_1$ and $\tilde{\Sigma}_2$ can be consistently estimated by

$$\tilde{\Sigma}_{1n} = n^{-1} \sum_{i=1}^{n} \int_0^\infty [Z_i - \bar{Z}(t; \tilde{\beta}_n)]^{\otimes 2} \lambda(\tilde{\beta}_n^T Z_i + \tilde{h}_n(t))\tilde{Y}_i(t)\Delta_i(t; \hat{\theta}_n) d\tilde{h}_n(t),$$

and

$$\tilde{\Sigma}_{2n} = n^{-1} \sum_{i=1}^{n} \int_0^\infty [Z_i - \bar{Z}(t; \tilde{\beta}_n)] Z_i^T \dot{\lambda}_\epsilon(\tilde{\beta}_n^T Z_i + \tilde{h}_n(t))\tilde{Y}_i(t)\Delta_i(t; \hat{\theta}_n) d\tilde{h}_n(t),$$

respectively, where

$$\bar{Z}(t; \tilde{\beta}_n) = \sum_{i=1}^{n} \frac{Z_i \lambda_\epsilon(\tilde{\beta}_n^T Z_i + \tilde{h}_n(t))\tilde{Y}_i(t)\Delta_i(t; \hat{\theta}_n)\tilde{B}_n(t, T_i)}{\sum_{i=1}^n \lambda(\tilde{\beta}_n^T Z_i + \tilde{h}_n(t))\tilde{Y}_i(t)\Delta_i(t; \hat{\theta}_n)},$$

$$\tilde{B}_n(t, s) = \exp\left(\int_s^t \frac{\sum_{i=1}^n \dot{\lambda}(\tilde{\beta}_n^T Z_i + \tilde{h}_n(x))\tilde{Y}_i(x)\Delta_i(x; \hat{\theta}_n)}{\sum_{i=1}^n \lambda(\tilde{\beta}_n^T Z_i + \tilde{h}_n(x))\tilde{Y}_i(x)\Delta_i(x; \hat{\theta}_n)} d\tilde{h}_n(x)\right).$$

Hence, a consistent estimator of $\Sigma_{\tilde{\beta}_n}$ is given by $\tilde{\Sigma}_{\tilde{\beta}_n} = \tilde{\Sigma}_{2n}^{-1}\tilde{\Sigma}_{1n}(\tilde{\Sigma}_{2n}^{-1})^T$.

3. Simulation study

A simulation study is conducted to compare the performance of the two estimators, $\hat{\beta}_n$ and $\tilde{\beta}_n$. We generated T following the proportional odds model with $h(t) = \log(t/10)$ and $\beta = (\beta_1 = 1, \beta_2 = 1)^T$. The resulting T has the survivorship function

$$P(T > t | Z_1, Z_2) = \frac{1}{1 + \exp\{\log(t/10) + Z_1 + Z_2\}},$$

where Z_1 is an ordinal variable with $P(Z_1 = i) = 0.25$ for $i = 1, 2, 3, 4$ and Z_2 is a Bernoulli random variable with probability 0.5. Note that under this set-up, the p^{th} percentile of T at (Z_1, Z_2) is $t_p = 10\exp\{\log((1-p)/p) - (Z_1 + Z_2)\}$, which decreases as Z_1 or Z_2 increases.

Case 1: $G(x; \theta_v) = 1 - e^{-\theta_v x}$

The truncation variable V was generated from an exponential distribution with mean equal to $\theta_v = 10, 2.0, 0.6$ such that the truncation probabilities are equal to 0.2, 0.5 and 0.7, respectively. Sample size is set at $n = 100, 200, 300$. The replication time is 1000. Using $\hat{\Sigma}_{\hat{\beta}_n}$ and $\tilde{\Sigma}_{\tilde{\beta}_n}$, we also calculated the estimated standard deviations of $\hat{\beta}_n = (\hat{\beta}_{1n}, \hat{\beta}_{2n})^T$ and $\tilde{\beta}_n = (\tilde{\beta}_{1n}, \tilde{\beta}_{2n})^T$, denoted by estd$(\hat{\beta}_n)$ and estd$(\tilde{\beta}_n)$, respectively. For $i = 1, 2$, an approximate 95% confidence interval for β is also constructed using $\hat{\beta}_{in} \pm z_{0.025}$estd$(\hat{\beta}_{in})$ (or $\tilde{\beta}_{in} \pm z_{0.025}$estd$(\tilde{\beta}_{in})$), where $z_{0.025}$ is the 0.025 upper percentile point of the standard normal distribution. Let $C(\hat{\beta}_{in})$ and $C(\tilde{\beta}_{in})$ denote the empirical coverage using the above procedure. Table 1 shows the simulated biases, simulated standard deviations (std), the estimated standard deviation (estd), $C(\hat{\beta}_{in})$, $C(\tilde{\beta}_{in})$ and the ratio of the simulated root mean squared error (rmse) of $\hat{\beta}_{in}$ to that of $\tilde{\beta}_{in}$ (denoted by ratio). Table 1 also shows the proportion of left-truncation $P(T < V)$ (denoted by q).

Case 2: $G(x; \theta) = 1 - e^{-(\theta_v x)^\gamma}$

The simulation set-up is the same as Case 1 except that the truncation variable V was generated from a Weibull distribution with scale parameter θ_v and shape parameter γ equal to $(\theta_v, \gamma) = (11.6, 2.0), (2.86, 2.0), (0.2, 0.46)$ such that the truncation probabilities are equal to 0.2, 0.5 and 0.7, respectively. The simulation results are listed in Table 2.

Table 1. simulated biases, std., estd and coverages of $\hat{\beta}_n$ and $\tilde{\beta}_n$ (Case 1)

			$\hat{\beta}_{1n}$				$\tilde{\beta}_{1n}$				
θ_v	q	n	bias	std	estd	$C(\hat{\beta}_{1n})$	bias	std	estd	$C(\tilde{\beta}_{1n})$	ratio
10	0.2	100	0.013	0.231	0.212	0.939	0.014	0.223	0.209	0.942	0.965
10	0.2	200	0.017	0.164	0.152	0.943	0.010	0.157	0.150	0.946	0.954
10	0.2	300	0.011	0.143	0.136	0.949	0.012	0.141	0.134	0.951	0.986
2.0	0.5	100	0.016	0.289	0.264	0.937	-0.014	0.263	0.248	0.941	0.909
2.0	0.5	200	0.018	0.199	0.188	0.941	0.013	0.175	0.167	0.942	0.878
2.0	0.5	300	0.012	0.166	0.158	0.946	0.008	0.154	0.147	0.949	0.926
0.6	0.7	100	-0.037	0.395	0.364	0.935	-0.031	0.316	0.299	0.939	0.800
0.6	0.7	200	0.011	0.318	0.297	0.941	0.016	0.239	0.226	0.945	0.753
0.6	0.7	300	0.019	0.215	0.205	0.945	0.004	0.173	0.200	0.948	0.801
			$\hat{\beta}_{2n}$				$\tilde{\beta}_{2n}$				
θ_v	q	n	bias	std	estd	$C(\hat{\beta}_{1n})$	bias	std	estd	$C(\tilde{\beta}_{1n})$	ratio
10	0.2	100	0.053	0.477	0.449	0.939	0.027	0.462	0.440	0.942	0.963
10	0.2	200	0.016	0.350	0.329	0.943	-0.007	0.337	0.322	0.944	0.962
10	0.2	300	-0.009	0.296	0.286	0.948	0.011	0.278	0.270	0.951	0.939
2.0	0.5	100	0.050	0.606	0.565	0.937	0.040	0.525	0.502	0.942	0.865
2.0	0.5	200	0.044	0.455	0.425	0.942	0.022	0.417	0.399	0.945	0.913
2.0	0.5	300	0.013	0.354	0.337	0.946	-0.015	0.325	0.312	0.949	0.918
0.6	0.7	100	-0.081	0.821	0.769	0.935	-0.085	0.674	0.641	0.939	0.823
0.6	0.7	200	0.063	0.602	0.570	0.942	0.054	0.540	0.517	0.942	0.896
0.6	0.7	300	0.049	0.433	0.417	0.944	0.021	0.374	0.361	0.947	0.830

Table 2. simulated biases, std., estd and coverages of $\hat{\beta}_n$ and $\tilde{\beta}_n$ (Case 2)

				$\hat{\beta}_{1n}$				$\tilde{\beta}_{1n}$				
θ_v	γ	q	n	bias	std	estd	$C(\hat{\beta}_{1n})$	bias	std	estd	$C(\tilde{\beta}_{1n})$	ratio
11.6	2.0	0.2	100	0.025	0.199	0.185	0.938	-0.030	0.195	0.186	0.940	0.983
11.6	2.0	0.2	200	0.014	0.174	0.166	0.942	-0.010	0.172	0.165	0.943	0.982
11.6	2.0	0.2	300	-0.013	0.150	0.145	0.947	-0.007	0.147	0.142	0.947	0.976
1.00	2.0	0.5	100	0.042	0.280	0.263	0.936	0.011	0.247	0.237	0.938	0.873
1.00	2.0	0.5	200	0.018	0.178	0.169	0.941	0.016	0.152	0.146	0.942	0.854
1.00	2.0	0.5	300	0.022	0.134	0.129	0.945	0.015	0.125	0.121	0.946	0.926
0.07	0.5	0.7	100	0.062	0.288	0.272	0.935	0.023	0.241	0.229	0.937	0.821
0.07	0.5	0.7	200	0.016	0.183	0.176	0.941	-0.003	0.154	0.146	0.942	0.792
0.07	0.5	0.7	300	0.003	0.157	0.151	0.945	0.008	0.129	0.124	0.945	0.823
				$\hat{\beta}_{2n}$				$\tilde{\beta}_{2n}$				
θ_v	γ	q	n	bias	std	estd	$C(\hat{\beta}_{2n})$	bias	std	estd	$C(\tilde{\beta}_{2n})$	ratio
11.6	2.0	0.2	100	0.063	0.490	0.457	0.939	-0.067	0.478	0.427	0.941	0.977
11.6	2.0	0.2	200	-0.048	0.344	0.331	0.943	-0.042	0.337	0.325	0.944	0.978
11.6	2.0	0.2	300	-0.038	0.288	0.280	0.947	-0.020	0.275	0.269	0.948	0.949
1.00	2.0	0.5	100	0.027	0.502	0.476	0.936	0.011	0.451	0.429	0.937	0.897
1.00	2.0	0.5	200	0.030	0.384	0.367	0.940	0.014	0.340	0.332	0.942	0.883
1.00	2.0	0.5	300	0.008	0.323	0.314	0.946	-0.005	0.288	0.275	0.946	0.889
0.07	0.5	0.7	100	0.007	0.559	0.528	0.935	0.006	0.457	0.433	0.937	0.817
0.07	0.5	0.7	200	0.016	0.364	0.346	0.940	0.014	0.292	0.282	0.941	0.802
0.07	0.5	0.7	300	0.010	0.308	0.295	0.945	-0.009	0.252	0.244	0.945	0.820

Based on the results of Tables 1 and 2, we have the following conclusions:

(1) The standard deviations of both estimators increase as the proportion of left-truncation q increases. When truncation is severe (i.e. $q = 0.7$) and $n = 100$, the biases of both estimators can be large. In terms of root mean squared error, the estimator $\tilde{\beta}_n$ outperforms $\hat{\beta}_n$ for all the cases considered. The improvement in using $\tilde{\beta}_n$ can be very significant when truncation is severe (i.e. $q = 0.7$). For case 1, the ratio of squared root mean squared error of $\tilde{\beta}_n$ to that of $\hat{\beta}_n$ ranges from 0.753 to 0.986. For case 2, the ratio of squared root mean squared error of $\tilde{\beta}_{2n}$ to that of $\hat{\beta}_{2n}$ ranges from 0.792 to 0.983.

(2) When $n = 100$, the estimated standard deviations are smaller that the empirical standard deviations, which makes the coverage of 95% confidence intervals smaller than the nominal

level. However, when $n = 300$, the coverage of 95% confidence intervals is close to the nominal level for all the cases considered.

4. Discussion

The semiparametric estimators proposed in this article are designed to incorporate both information contained in the data and the available information on the truncation distribution, and are expected to have better performance than nonparametric methods. Our simulation study indicates that under the additional assumption $G(x) = G(x;\theta)$, the estimators $\tilde{\beta}_n$ can perform better than the estimators $\hat{\beta}_n$. In practice the validity of the assumption $V \sim G(x;\theta)$ can be checked by plotting $\hat{G}_n(x)$ against $\hat{G}(x;\hat{\theta}_n)$, where $\hat{G}_n(x)$ is the NPMLE of $G(x)$ (see Wang (1987)) and given by

$$\hat{G}_n(x) = \left[\sum_{i=1}^{n} \frac{1}{1-\hat{F}_n(V_i-)}\right]^{-1} \sum_{i=1}^{n} \frac{I_{[V_i \le x]}}{1-\hat{F}_n(V_i-)},$$

where

$$\hat{F}_n(x) = 1 - \prod_{u \le x}\Big[1 - \frac{N_F(du)}{Y(u)}\Big].$$

where $Y(u) = \sum_{i=1}^{n} Y_i(u)$, $N_F(u) = \sum_{i=1}^{n_1} N_i(u)$ and $N_F(du) = N_F(u) - N_F(u-)$. Besides, for large samples, Wang (1991) studied the properties of the NPMLE $\hat{G}_n(x)$ and showed that its asymptotic distribution is Gaussian. This can be used for testing the null hypothesis $H_0 : G(x) = G(x;\theta_0)$. Let $I_1 = [V_{(1)} \equiv a_0, a_1]$, $I_j = (a_{j-1}, a_j]$ $(j = 1, \ldots, M-1)$ and $I_M = (a_{M-1}, a_M \equiv V_{(n)}]$ be a partition of the interval $[V_{(1)}, V_{(n)}]$, where $V_{(1)}$ and $V_{(n)}$ denote the smallest and largest observation of V_i's. For $j = 1, \ldots, M$, let $\hat{G}_n(I_j) = \hat{G}_n(a_j) - \hat{G}_n(a_{j-1})$ and $G(I_j;\theta_0) = G(a_j;\theta_0) - G(a_{j-1};\theta_0)$. Then $\eta = \sqrt{n}\{\hat{G}_n(I_1) - G(I_1;\theta_0), \ldots, \hat{G}_n(I_M) - G(I_M;\theta_0)\}$ has an asymptotic mean zero normal distribution with covariance matrix Σ (see Theorem 4.1 of Wang (1991)). Thus under H_0, the statistic $\hat{T}_W = \eta^T \hat{\Sigma}^{-1} \eta$ has an asymptotic chi-square distribution with $M-1$ degrees of freedom, where $\hat{\Sigma}$ is a consistent estimator for Σ.

It requires further research to extend our approach to left-truncated and right-censored data.

References

Asgharian, M., M'Lan, C. E., and Wolfson, D. B. (2002). Length-biased sampling with right

censoring: an unconditional approach. *Journal of the American Statistical Association* **97**, No. 457, 201-209.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statist. Med.*, **2**, 273-277.

Chen, K., Jin, Z and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, **89**, 659-668.

Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, **82**, 4, 835-845.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. of Roy. Statist. Soc. Ser.B*, **34**, 187-220.

Fleming and Harrington, (1991). Counting Process and Survial Analysis .*John Wiley and Sons*.

Fiegelson, E. D. and Babu, G. J. (Eds.) (1992). Statistical Challenge in Modern Astronomy. Springer, New York.

Keiding, N., Bayer, T., and Watt-Boolsen, S. (1987). Confirmatory analysis using left truncation of the life times of primary survivors. *Statistics in Medicine*, **6**, 939-944.

Lagakos, S. W., Barraj, L. M. and DeGruttola, V. (1988). Nonparametric analysis of truncated survival data with application to AIDS. *Biometrika*, **75**, 515-523.

Murphy, S. A., Rossini, A. J. and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *J. Am. Statist. Assoc.*, **92**, 968-976.

Shen, P.-S. (2012). Semiparametric analysis of transformation models with left-truncated and right-censored data. *Computational Statistics*, **26**, 521-537.

Wang, M.-C., (1987), Product-limit estimates: a generalized maximum likelihood study. *Communications in Statistics-Theory and Methods*, **6**, 3117-3132.

Wang, M.-C. (1989). A semiparametric model for randomly truncated data. *Journal of the*

American Statistical Association **84**, 742-748.

Wang, M.-C. 1991, Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, *86*, 130-143.

Wang, M.-C. (1996). Hazard regression analysis with length-biased data. *Biometrika*, **83**, 343-354.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Annals of Statistics*, **13**, 163-177.

Ying, S and Prentice, R. (1999). Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Association*, **94**, 125-136.